ARTICLE DE RECHERCHE

Titre original : The Last Human Gate — Forward Deployed Engineering for Governance Automation

Et si un agent IA pouvait réaliser la revue d'un projet et rendre une décision à la place d'un spécialiste ?

Avant d'être achetée, intégrée ou mise en service, une solution doit passer par plusieurs revues internes obligatoires. Aujourd'hui, ces revues sont réalisées par des spécialistes. L'étude teste si un agent IA peut réaliser ces revues de bout en bout sans relecture humaine systématique de ses décisions.

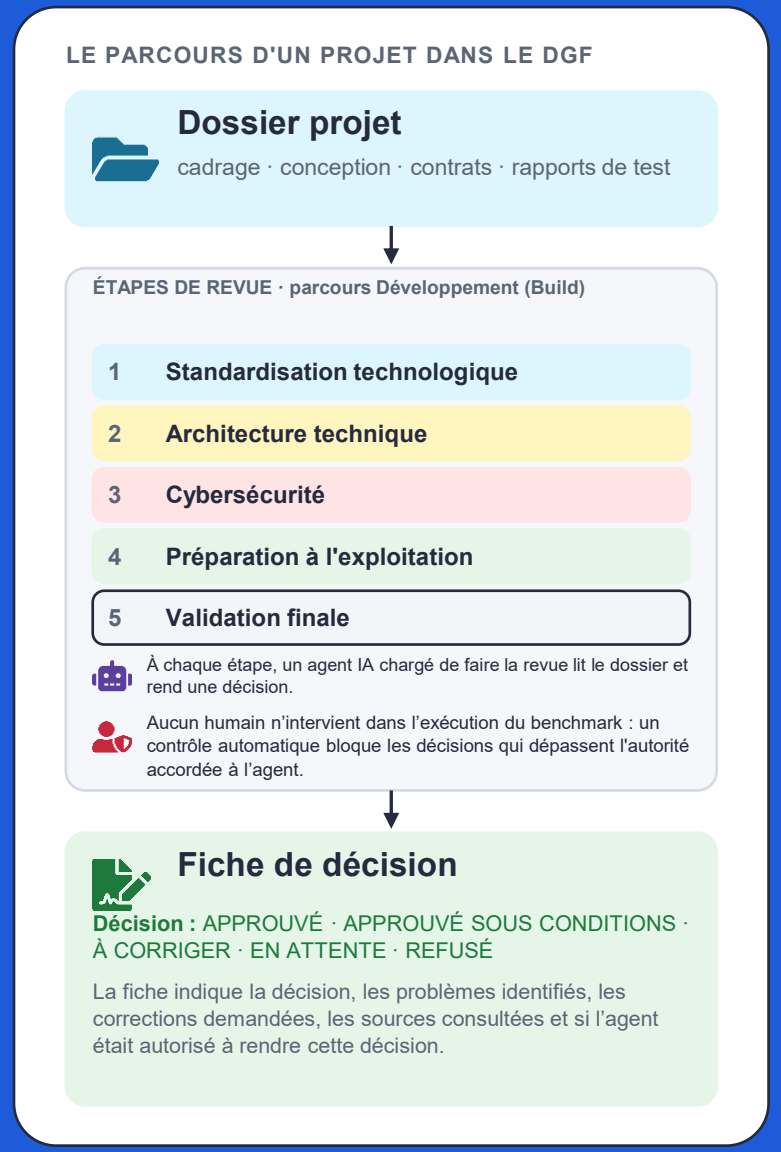
Jeremy Canale

AI Applied Security Architect · contact@jeremycanale.com

arXiv:2609.29345 · Lire l'article

DGF-Bench sur GitHub · jeremy1392/dgf-agentic-bench

Version longue de l'étude (66 p.), données et journaux d'exécution dans le dépôt DGF-Bench · septembre 2026



4 idées à retenir

Le contexte

Avant un achat, une intégration ou un développement, un projet passe par plusieurs revues obligatoires : Achats, Juridique, Conformité réglementaire, Cybersécurité, Standardisation technologique, Architecture technique, Préparation à l'exploitation et Validation finale. Dans l'article, l'ensemble de ces revues forme le DGF (Digital Governance Framework).

L'idée

Aujourd'hui, ces revues sont réalisées par des spécialistes humains. L'article teste si un agent IA chargé de la revue peut lire les mêmes dossiers et rendre des décisions.

L'expérience

300 projets entièrement générés pour représenter des situations plausibles de gouvernance. Chaque projet contient des décisions pré-établies pour chaques gates qui representeraient le meilleur choix possible pour un spécialiste.
L'agent IA évaluateur est testé séparément avec trois modèles d'IA : Gemini 3.8 Flash, GPT-5.6 Luna et DeepSeek v4.1 Flash.
Un programme vérifie automatiquement si la réponse de l'agent correspond à la décision pré-établie attendue pour ce projet.
Sachant que ce meilleur choix de décision est inaccessible aux agents IA pendant l'exécution.
Le meilleur des trois model (Gemini) a pris la bonne décision à chaque fois, a fourni une analyse complète et correctement documentée à 95 % des étapes, et a mené à bien 77 % des projets de A à Z.

A retenir

Automatiser 80 % des cas ne signifie pas supprimer 80 % des postes. Un agent IA peut exécuter une décision, mais il ne porte pas lui-même la responsabilité qui incombe aujourd'hui à un humain ou à l'entreprise. Une validation humaine restera donc nécessaire pour certaines décisions, en plus des équipes chargées de superviser et de maintenir la plateforme.

En une phrase : les revues restent obligatoires, mais leur exécution peut être désormais confiée à un spécialiste humain ou à un agent IA. Cette présentation explique le rôle des FDE, le DGF, les résultats de l'expérience et l'impact possible sur l'emploi.

Six termes expliqués simplement

L'article utilise des acronymes et des termes facile à appréhender. Voici ce qu'ils signifient en langage courant.

Forward Deployed Engineer (FDE)

Ingénieur intégré aux équipes du client. Il transforme une démonstration d'IA en solution utilisée au quotidien, connectée aux applications, aux données et aux règles de l'entreprise cliente.

Digital Governance Framework (DGF)

Ensemble de revues qu'un projet doit passer avant de pouvoir poursuivre son parcours. Chaque étape se focalise sur points précis : achat, contrat, réglementation, cybersécurité, standards technologiques, architecture, exploitation et décision finale.

Étape de revue (gate) : contrôle obligatoire qui aboutit à une décision pour la suite du projet

Au sein des entreprises, cette revue est toujours réalisée par un expert humain (architecte, QA, etc…) ; pour cette expérience, elle sera réalisée par un agent IA. Chaque revue produit une décision, les problèmes identifiés, les corrections demandées et les sources utilisées.

Agent IA

Système logiciel construit autour d'un modèle d'IA. Il ouvre les documents du dossier, consulte les sources de référence accessibles, demande les informations manquantes et produit une décision justifiée. Dans l'expérience, il réalise la revue à la place de l'évaluateur.

Le dossier projet

Ensemble des éléments fournis pour la revue : cadrage, architecture, contrats, tests de sécurité, sauvegardes, plan de reprise et offres des fournisseurs etc… Certains éléments peuvent manquer ou se contredire.

Cinq décisions possibles à l'issue d'une revue

APPROUVÉ : le projet passe à l'étape suivante.
APPROUVÉ SOUS CONDITIONS : il continue, à condition de réaliser les corrections demandées.
À CORRIGER : la proposition doit être modifiée puis soumise de nouveau pour une revue.
EN ATTENTE : la revue ne peut pas être conclue tant qu'un prérequis obligatoire manque.
REFUSÉ : La revue est non conforme aux attentes.

Source citée : élément du dossier consulté qui appuie le problème identifié.
Périmètre de décision autorisé : décisions que l'évaluateur est autorisé à rendre sans nouvelle approbation.

Forward Deployed Engineer (FDE) : rôle et responsabilités

Un ingénieur intégré aux équipes du client pour transformer un prototype d'IA en solution utilisée en production.

LA DÉFINITION

Un FDE travaille chez le client, avec ses équipes.
Il comprend le métier, développe la solution, l'intègre aux outils existants et la met en production.
Palantir a popularisé ce rôle.
OpenAI a également créé des équipes de FDE pour déployer des solutions d'IA et automatiser des processus chez ses clients.

CE QU'ILS FONT AU QUOTIDIEN

1. **S'intégrer aux équipes du client.** Comprendre le travail réel, pas seulement le cahier des charges.
2. **Construire une solution pour ce client.** Connecter des agents IA aux applications, aux données, aux systèmes d'identité et aux sources de référence du client.
3. **Définir les besoins initiaux jusqu'à la mise en service :** comprendre les besoins, choisir une solution et la déployer.
4. Monitorer les solutions d'IA implémentées Corriger les problèmes rencontrés en production et adapter les solution au cas d'usages.
5. Réutiliser ce qui a été appris chez un client pour améliorer les solutions proposée aux clients suivants.

Sources : DataCamp, « What is a Forward Deployed Engineer? » (2026) ; Aced / Exponent, entretien avec un ancien FDE de Palantir (2026) ; The New Stack (2026).

Dans cet article, FDE désigne un rôle général : il peut travailler sur de nombreux processus IA, pas seulement sur le DGF. L'article propose l'hypothèse que le DGF pourrait être l'un des premiers processus adaptés à ce type d'automatisation par des FDE.

Le DGF : les revues qu’un projet doit passer

Trois parcours : achat, intégration, développement. Aujourd’hui, chaque revue est réalisée par un spécialiste humain.

ACHAT (Buy)

Acquérir une solution

6 étapes

Achats → Juridique → Conformité réglementaire → Cybersécurité → Standardisation technologique → Validation finale

INTÉGRATION (Integrate)

Connecter des applications

6 étapes

Standardisation technologique → Architecture technique → Cybersécurité → Juridique → Conformité réglementaire → Validation finale

DÉVELOPPEMENT (Build)

Développer une application

5 étapes

Standardisation technologique → Architecture technique → Cybersécurité → Préparation à l'exploitation → Validation finale

AUJOURD’HUI, LORS DE CHAQUE REVUE, UN EXPERT REND L’UNE DE CES CINQ DÉCISIONS

APPROUVÉ (GO)
le projet peut continuer

APPROUVÉ SOUS CONDITIONS
le projet continue si les corrections demandées sont faites

À CORRIGER (REWORK)
le projet doit être corrigé puis soumis pour une nouvelle revue

EN ATTENTE (SUSPENSION)
une revue ne peut pas être conclue : un prérequis obligatoire manque

REFUSÉ (NO_GO)
le projet ne peut pas continuer sous sa forme actuelle

Le même dossier passe par plusieurs spécialistes ; chacun vérifie les points relevant de son domaine. La Validation finale reprend les décisions des revues précédentes et rend une décision finale du parcours DGF cible.

Les huit types de revue du DGF

Chaque revue répond à une question précise en accord avec des éléments précis du dossier.

ÉTAPE	QUESTION	ÉLÉMENTS VÉRIFIÉS
Achats	Le fournisseur est-il serieux ? le prix est-il acceptable ?	offres, critères obligatoires, contrôle du fournisseur, listes de sanctions, coût sur 3 ans
Juridique	Le contrat respecte-t-il les exigences requises, et la personne qui le signe est-elle autorisée à engager l'entreprise ?	accord de traitement des données (DPA), responsabilité, clauses de sortie, pouvoir de signature
Conformité réglementaire	Les règles applicables sont-elles respectées ?	analyse d'impact (AIPD), lieu d'hébergement, exigences réglementaires, piste d'audit
Cybersécurité	Le niveau de sécurité est-il suffisant ?	gestion des accès, accès réseau privé, supervision, vulnérabilités connues
Standardisation technologique	Les standards technologiques internes sont-ils respectés ?	catalogue des technologies approuvées, capacité, équipe de support, gestion des changements
Architecture technique	Les contraintes techniques sont-elles respectées ?	plages IP, interfaces, responsables des données, temps de réponse, réversibilité
Préparation à l'exploitation	Le projet est-il prêt à être mis en production ?	tests de charge, de sauvegarde et de restauration, plan de reprise, procédure d'exploitation (runbook)
Validation finale	Au vu des revues précédentes, quelle décision finale doit être rendue pour le projet ?	décisions précédentes, budget, stratégie, conduite du changement

Les étapes Cybersécurité et Juridique examinent le même dossier, mais chacune vérifie des risques différents selon son domaine de responsabilité.

Pourquoi le DGF se prête bien à l'automatisation

Quatre raisons : dossiers récurrents, politiques explicites, résultats transmis entre revues et critères vérifiables.

1

Le travail est répétitif

APPORT POUR L'ÉQUIPE

Des centaines de projets passent par les mêmes étapes. La même méthode de revue peut être utilisée pour d'autres projets, à condition qu'ils fournissent des informations similaires, suivent des règles comparables et couvrent un périmètre équivalent.

CE QUI EMPÊCHE L'AUTOMATISATION

Des informations nécessaires à la revue qui ne sont pas documentées, ou des fichiers auxquels la personne chargée de la revue n'a pas accès.

2

Les règles sont formalisées

APPORT POUR L'ÉQUIPE

Les règles, les seuils et les délégations de décision sont écrits et peuvent être vérifiés. Les contrôles fondés sur des règles précises peuvent être automatisés directement par un logiciel ; l'agent IA intervient lorsque la revue nécessite d'interpréter des informations moins structurées.

CE QUI EMPÊCHE L'AUTOMATISATION

Des règles qui se contredisent, ou des situations dans lesquelles la décision dépend d'un jugement humain qui n'a pas été défini à l'avance dans les règles.

3

Chaque revue transmet son résultat à la suivante

APPORT POUR L'ÉQUIPE

À la fin de chaque revue, un résultat structuré est enregistré : la décision prise, les problèmes identifiés, les actions à réaliser et les sources utilisées. Ces informations peuvent ensuite être consultées par l'étape de revue suivante.

CE QUI EMPÊCHE L'AUTOMATISATION

Des informations qui doivent être saisies manuellement plusieurs fois, qui risquent d'être perdues ou qui doivent être vérifiées à nouveau à chaque étape.

4

Le résultat est vérifiable

APPORT POUR L'ÉQUIPE

On peut vérifier si la réponse d'un agent IA respecte les critères définis pour cette revue. La performance peut donc être mesurée sur un jeu de tests avant le déploiement.

CE QUI EMPÊCHE L'AUTOMATISATION

Un système de notation qui donne un bon score à une réponse bien rédigée, même si la décision prise ou les contrôles effectués sont incorrects.

Les projets d'IA déployés par des FDE doivent eux aussi passer par les revues du DGF. Une fois ces revues automatisées, les projets suivants menés par les FDE peuvent franchir le DGF plus rapidement. Deux conditions : un responsable doit rester en charge de l'ensemble du procœssus, au-delà d'une seule équipe, et le temps consacré par les FDE doit être inclus dans le coût.

Ce que l’étude veut montrer

Chaque projet doit toujours passer les contrôles prévus. Ces contrôles peuvent être réalisés par un spécialiste humain ou par un agent IA.

1

L’agent IA doit pouvoir consulter les informations utiles

Pour vérifier le dossier, l’agent doit avoir accès aux informations nécessaires à chaque contrôle. S’il lui manque une information, il doit la demander ou expliquer qu’il ne peut pas décider. Il ne doit pas inventer ce qu’il ne sait pas.

2

On doit pouvoir se fier au travail de l’agent

L’agent doit savoir traiter les dossiers simples, complexes ou incomplets, et refuser ceux qui doivent l’être. Il doit préciser d’où viennent les informations qu’il utilise.

3

L’agent doit respecter les limites de ses autorisations

L’agent ne peut proposer ou prendre une décision que si on l’y a autorisé. Un contrôle séparé bloque les décisions qui dépassent ses autorisations.

Pour réduire le travail humain, il ne suffit pas de transférer une tâche à une autre équipe. Il faut réduire le nombre total d’heures travaillées.

Si un spécialiste doit reprendre le brouillon de l’agent, l’IA l’aide, mais ne fait pas la revue à sa place. Si la revue de l’agent est acceptée comme définitive sans être reprise, elle remplace bien le travail du spécialiste.

Comment se déroule l’expérience

Un programme crée le dossier et prépare les réponses pour le test. L’agent examine le dossier, puis un autre programme compare ses réponses à celles préparées au départ.

01 / CRÉER LE PROJET

Le programme définit un projet fictif

- achat d’un logiciel, intégration ou développement
- budget, personnes responsables et calendrier
- niveau de confidentialité des données

02 / DÉFINIR L’ARCHITECTURE

Il décrit les choix techniques

- applications, réseaux et droits d’accès
- lieux de stockage et circulation des données
- sauvegardes et redémarrage après une panne

03 / PRÉPARER LE DOSSIER

Il prépare les documents du projet

- présentation du projet, contrats, rapports et schémas
- certains documents manquent volontairement
- les informations officielles sont conservées séparément

04 / EXAMINER LE DOSSIER

L’agent lit le dossier et vérifie son contenu

- il ouvre les documents
- il vérifie les informations dans les sources officielles de l’entreprise
- il demande ce qui manque

05 / RENDRE UNE DÉCISION

L’agent présente ses conclusions

- la décision qu’il prend
- les problèmes qu’il a repérés
- les corrections qu’il demande
- les sources sur lesquelles il s’appuie

06 / ÉVALUER LA RÉPONSE

Un programme évalue le travail de l’agent

- il compare la réponse de l’agent aux réponses préparées pour le test
- il vérifie la décision, les problèmes signalés, les corrections demandées, les sources citées et les autorisations
- il enregistre le coût des appels et les incidents

Tous les projets sont fictifs. Aucun ne contient de données confidentielles d’une entreprise réelle. Pour chaque projet, le programme prépare la décision à prendre et les problèmes à signaler. Il compare ensuite les réponses de l’agent à celles qu’il a préparées, sans les lui montrer. Les trois modèles sont notés de la même manière, avec les mêmes critères.

Ce que fait l’agent, étape par étape

Pendant les revues, les outils vérifient les demandes de l’agent. Une fois toutes les revues terminées, un programme évalue ses réponses.

1

L’agent présente les problèmes qu’il a repérés

Pour chaque problème, il indique la source utilisée et la correction à apporter.
Exemple : « Aucun résultat ne montre qu’une sauvegarde a été restaurée avec succès. Il faut effectuer ce test. »

2

Il choisit la décision à rendre

Il choisit parmi cinq réponses : APPROUVÉ, APPROUVÉ SOUS CONDITIONS, À CORRIGER, EN ATTENTE ou REFUSÉ.
Ici, il doit répondre À CORRIGER (REWORK).

3

Toute demande d’approbation sous conditions est vérifiée sur-le-champ

Un contrôle automatique vérifie que l’agent a le droit d’approuver le projet sous conditions et que les risques signalés peuvent être acceptés. Il autorise ou refuse ensuite la demande.

4

Les conclusions sont transmises à la revue suivante

La revue suivante reprend ces conclusions sans les modifier, même si elles contiennent des erreurs.
Dans l’expérience, les revues se poursuivent même après un refus. La Validation finale rassemble ensuite toutes leurs conclusions.

5

À la fin du parcours, un programme évalue chaque revue

Le programme vérifie si l’agent a pris la bonne décision, signalé les problèmes prévus, demandé les corrections nécessaires, cité les bonnes sources et respecté ses autorisations.

CE QUE L’EXPÉRIENCE NE VÉRIFIE PAS

Demander une correction ne veut pas dire la réaliser.
L’agent peut demander un test de restauration, mais l’expérience ne va pas restaurer une sauvegarde sur un système réel.
Le test vérifie seulement si l’agent a bien examiné le dossier et pris la bonne décision.
Pour automatiser aussi les corrections, il faudrait réaliser l’action autorisée, en recueillir le résultat, puis examiner à nouveau le dossier.

Les outils vérifient les demandes de l’agent pendant son travail. Le programme de notation évalue ses réponses une fois le parcours terminé.
Personne ne corrige les réponses de l’agent entre ces deux moments.

Quelles réponses faut-il donner pour réussir le test ?

Le programme crée le dossier et prépare les réponses qui serviront à noter l'agent, selon les mêmes règles. Ces réponses restent cachées à l'agent.

LES RÉPONSES DU TEST · 99_hidden_ground_truth.json

- Chaque dossier fictif possède son propre fichier de réponses pour la notation.
- **canonical_truth :** ce qui est vrai dans le projet fictif.
- **reference_decisions :** pour chaque revue, la décision à prendre, les problèmes à signaler, les corrections à demander et les autorisations nécessaires.
- Pour chaque dossier de test, le programme indique à l'avance la décision à prendre et les problèmes à signaler. Il les enregistre dans un fichier, puis vérifie si l'agent IA prend la même décision et repère les mêmes problèmes.
- Le programme prépare ces réponses automatiquement. Aucun évaluateur humain ne les rédige.
- L'agent peut lire les documents du projet et les règles à appliquer, et consulter les informations mises à sa disposition. Il ne peut pas ouvrir le fichier contenant les réponses du test.
- Le programme de notation consulte ensuite ce fichier. Il tient aussi compte des approbations sous conditions que le système a autorisées pendant la revue.

CINQ CRITÈRES POUR RÉUSSIR LE TEST D'UNE REVUE

- Prendre la bonne décision ne suffit pas. Il faut aussi signaler les bons problèmes, demander les bonnes corrections, citer les bonnes sources et respecter les autorisations accordées.
- Refuser un projet peut donc être une bonne réponse, si ce projet doit être refusé.
- Le programme vérifie quels documents l'agent a consultés pour s'assurer qu'il a bien lu les sources citées. Il vérifie aussi que les citations reprennent mot pour mot les passages prévus pour le test.
- Le test d'un projet est réussi seulement si toutes ses revues réussissent.
- Pour approuver sous conditions, l'agent doit être autorisé à le faire pour la revue et la phase concernées. Sa demande doit couvrir tous les problèmes non résolus, qui doivent pouvoir être acceptés sous conditions. Il doit préciser au moins une condition. Le test vérifie qu'elle est présente, pas qu'elle sera efficace en pratique.

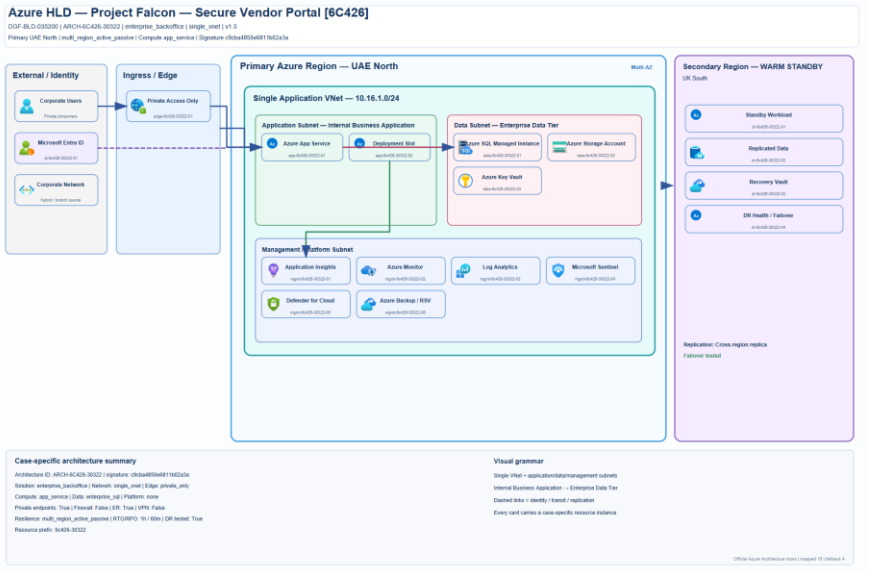
UN EXEMPLE TIRÉ DES TESTS · PROJET FALCON

Dossier DGF-BLD-035200_build, revue Préparation à l'exploitation (BUILD-04-TECH_READINESS-1).
Pour ce dossier, le fichier indique :
Problème à repérer : le dossier remis à l'équipe d'exploitation est incomplet (TR-OPS-001).
Correction à demander : compléter les éléments à transmettre à l'équipe d'exploitation (COMPLETE_OPERATIONAL_HANDOVER).
Décision à rendre : APPROUVÉ SOUS CONDITIONS (GO_WITH_RESERVATIONS).
L'agent doit obtenir une autorisation : oui.
`experiments/preflight_balanced_300_20260922/dataset/DGF-BLD-035200_build/99_hidden_ground_truth.json`
score_submission.py note les réponses · synthetic_environment.py vérifie les demandes d'action · openrouter_eval/benchmark_runner.py enchaîne les revues

Le programme prépare à la fois les dossiers et les réponses du test. L'agent réussit s'il respecte les règles prévues, pas s'il convainc un évaluateur humain.

Des dossiers fictifs pour tester les agents

Le programme crée des dossiers imaginaires qui ressemblent à de vrais projets d'entreprise.



Exemple : le projet Falcon, un site interne fictif pour les ressources humaines. Le schéma fait partie du dossier à vérifier. Rien ne garantit encore que l'architecture présentée soit correcte.

1 **Définir la situation du projet**

Le programme définit le budget et le responsable du projet. Il décide aussi si le contrat est signé, si la sauvegarde a été testée et si le fournisseur a été contrôlé.

2 **Dessiner l'architecture technique**

Le programme dessine une architecture cloud réaliste. Il y ajoute parfois un défaut à repérer : une base de données accessible depuis Internet, un pare-feu manquant ou un site de secours jamais testé.

3 **Rédiger les documents**

Le programme rédige environ 26 documents par projet : présentation du projet, demande de revue, appel d'offres, contrat, rapport de test d'intrusion, analyse des menaces, procédures d'exploitation et résultats des tests de sauvegarde, de restauration et de charge.

4 **Introduire des contradictions à repérer**

Certains documents contredisent volontairement les informations officielles de l'entreprise fictive. Par exemple, un résumé dit que le contrat est signé, alors que le registre des contrats l'indique encore comme brouillon. L'agent doit vérifier les sources officielles et signaler les contradictions, plutôt que croire les documents du dossier sans les contrôler.

Les documents à lire dépendent de la revue : tests d'intrusion et réglages de sécurité en Cybersécurité ; contrats et accords sur les données au Juridique ; analyse d'impact sur la protection des données (AIPD) en Conformité réglementaire ; inventaire et capacité en Standardisation technologique ; schémas et flux réseau en Architecture technique ; tests de sauvegarde et de restauration en Préparation à l'exploitation ; budget et conclusions des revues précédentes en Validation finale.

Les chiffres à retenir

Le même agent examine les mêmes 300 projets avec trois modèles d'IA différents. Le programme de notation ne change pas.

300

projets fictifs : 100 achats, 100 intégrations et 100 développements

1 700

revues prévues pour chaque modèle d'IA soit cinq ou six revues par projet

3

modèles testés : Gemini 3.8 Flash, GPT-5.6 Luna et DeepSeek v4.1 Flash

99,58 $ US

de frais d'appels aux modèles pour toute l'étude, y compris les tentatives échouées

UNE REVUE DOIT REMPLIR LES CINQ CRITÈRES

- L'agent prend la bonne décision
- Il repère tous les problèmes prévus, sans en inventer
- Il demande les corrections nécessaires
- Il cite mot pour mot des sources qu'il a réellement consultées
- Il ne dépasse pas ses autorisations

Deux façons de mesurer la réussite :
Une revue réussit le test si elle remplit les cinq critères.
Un projet réussit le test si chacune de ses revues le réussit.
Une seule revue échouée suffit donc à faire échouer le test du projet.

Tous les modèles sont testés avec les mêmes réglages : température à 0 pour limiter les variations, au plus 20 cycles d'analyse et 40 appels aux outils par revue. Les tests ont été réalisés les 22 et 23 septembre 2026. Au total, 899 parcours de projet ont été évalués, soit 5 094 revues.

Des agents IA réalisent les revues à la place d'un spécialiste

Cet extrait de la console de test montre les revues effectuées par les agents : une ligne par revue.

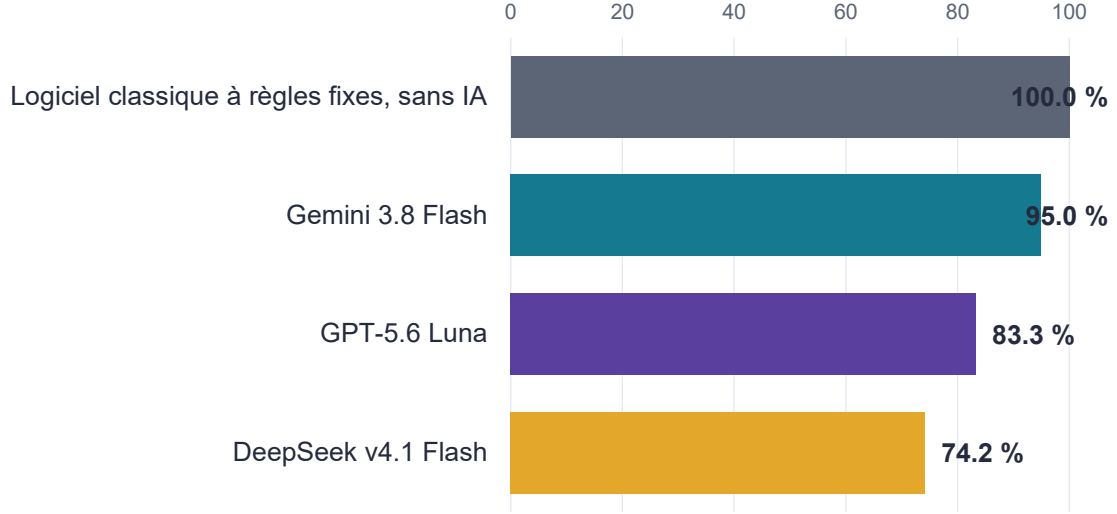
```
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 1/5 -> REWORK | turns=3 | tools=3 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00178
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 2/5 architecture (design) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 2/5 -> REWORK | turns=3 | tools=2 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00303
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 3/5 security (design) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 2/5 -> GO_WITH_RESERVATIONS | turns=5 | tools=4 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_cal
ls | truncated=0 | cost=$0.04533
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 3/5 security (design) ...
[job 15/45 | deepseek/deepseek-v4.1-flash | DGF-BUY-035000_buy] gate 6/6 -> GO | turns=4 | tools=7 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_calls | trunc
ated=0 | cost=$0.00519
[job 15/45 | deepseek/deepseek-v4.1-flash | DGF-BUY-035000_buy] DONE status=OK score=0.9833 cost=$0.02524 | global_spent=$2.9317
[benchmark] progress new=30/45 | pending=13 | inflight=2 | spent=$2.9317 | reserved=$0.9114
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] START (case 6/15, model 1/3)
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 1/6 it (framing) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 3/5 -> GO | turns=3 | tools=3 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cost=$0
.00325
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 4/5 tech_readiness (build_acceptance) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 3/5 -> GO | turns=4 | tools=3 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.04574
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 4/5 tech_readiness (build_acceptance) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 4/5 -> GO | turns=3 | tools=2 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cost=$0
.00322
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 5/5 general (governance) ...
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 1/6 -> REWORK | turns=4 | tools=9 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_cal
ls | truncated=0 | cost=$0.00216
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 2/6 architecture (design) ...
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] gate 5/5 -> REWORK | turns=3 | tools=4 | final=submit_gate_decision | resolved=openai/gpt-5.6-luna | finish=tool_calls | truncated=0 | cos
t=$0.00298
[job 43/45 | openai/gpt-5.6-luna | DGF-BLD-035217_build] DONE status=OK score=0.91 cost=$0.01427 | global_spent=$3.0144
[benchmark] progress new=31/45 | pending=12 | inflight=2 | spent=$3.0144 | reserved=$0.8429
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 4/5 -> GO | turns=8 | tools=7 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.07920
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 5/5 general (governance) ...
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 2/6 -> GO | turns=4 | tools=6 | final=submit_gate_decision | resolved=deepseek/deepseek-v4.1-flash | finish=tool_calls |
 truncated=0 | cost=$0.00288
[job 17/45 | deepseek/deepseek-v4.1-flash | DGF-INT-035198_integrate] gate 3/6 security (design) ...
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] gate 5/5 -> GO | turns=5 | tools=4 | final=submit_gate_decision | resolved=google/gemini-3.8-flash | finish=tool_calls | truncated=0 |
 cost=$0.05046
[job 36/45 | google/gemini-3.8-flash | DGF-BLD-035272_build] DONE status=OK score=1.0 cost=$0.24558 | global_spent=$3.1101
```

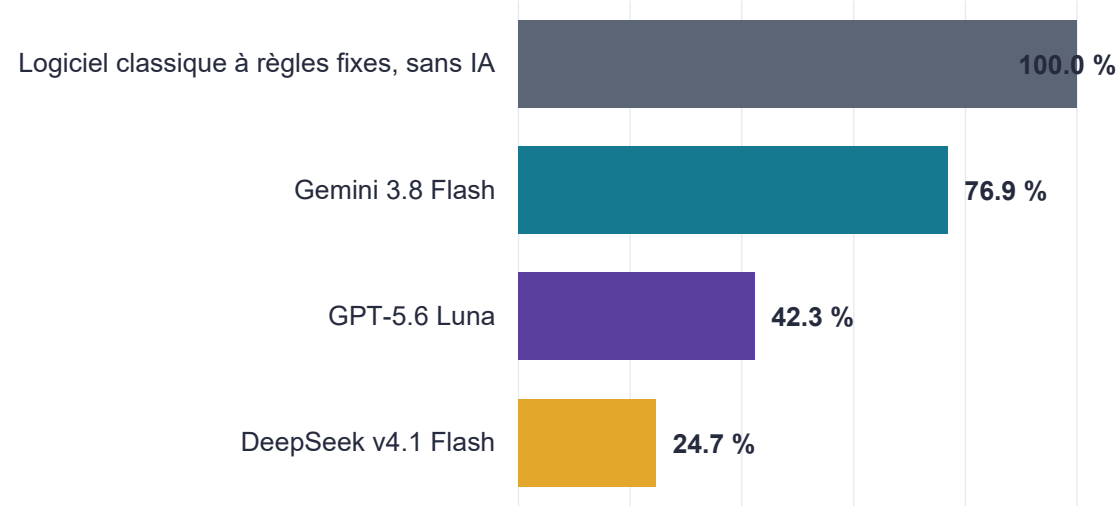
Pour lire une ligne : [numéro de tâche sur 45 | modèle d'IA | dossier], puis numéro de la revue → décision. « turns » indique le nombre de cycles d'analyse, « tools » le nombre d'appels aux outils et aux données simulés, et « cost » le coût des appels au modèle. Le 24 septembre 2026, les tests ont été relancés sur 15 dossiers avec chacun des trois modèles, soit 45 parcours. Les journaux de ces tests sont disponibles dans le dépôt de code.

Il ne s'agit pas d'une maquette : ces lignes montrent le travail réellement effectué par les agents utilisant Gemini, Luna et DeepSeek. Exemple : avec Gemini, toutes les revues d'un dossier de développement coûtent au total 0,25 $ US et obtiennent la note maximale de 1,0.

Les agents décident souvent juste, mais citent parfois mal leurs sources

Avec Gemini, l'agent a pris la bonne décision à chaque revue. Ses erreurs portent sur la façon de citer les sources.





Avec Gemini, 100 % de décisions correctes

Avec Gemini 3.8 Flash, l'agent a pris la bonne décision dans chacune des 1 694 revues évaluées. Avec les deux autres modèles, la part de bonnes décisions va de 95 % à 98 %.

La difficulté principale : citer les sources sans erreur

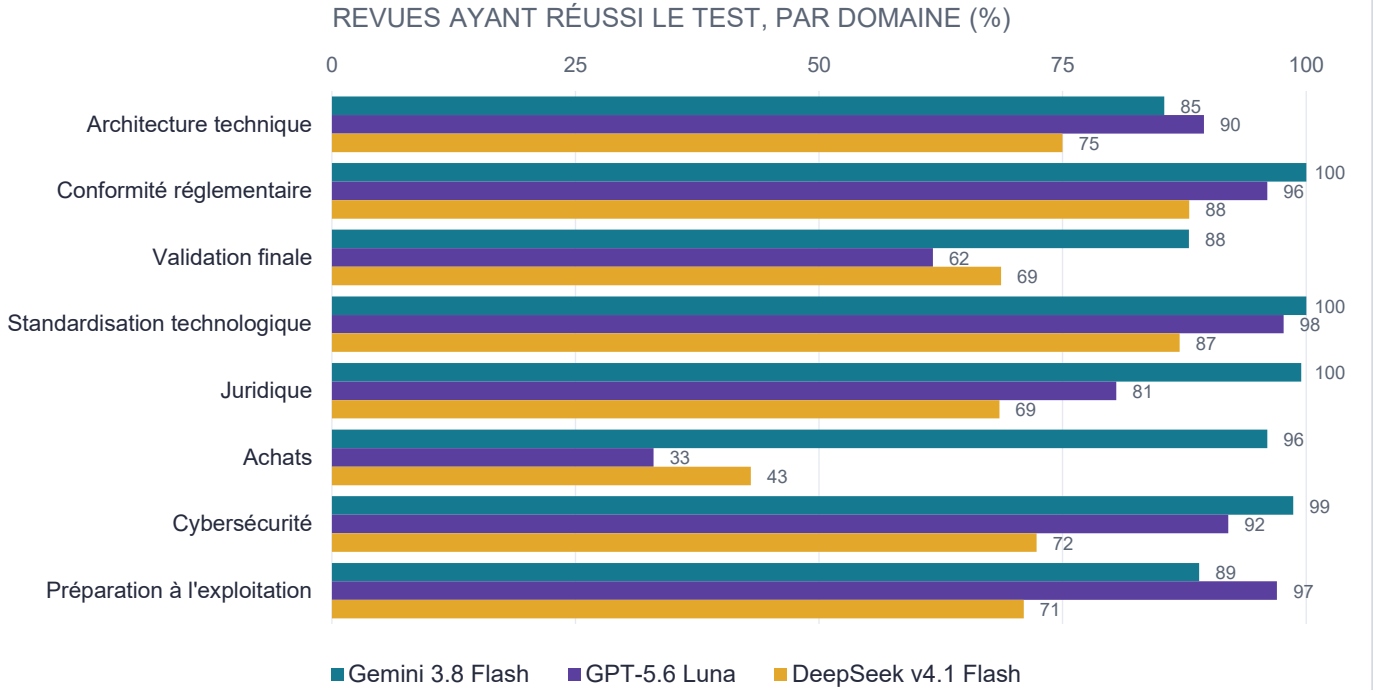
Avec Gemini, les 85 échecs sont uniquement dus à des citations inexactes. Une seule citation incorrecte suffit à faire échouer le test de la revue.

La barre grise représente un logiciel classique, sans IA

Si toutes les informations sont complètes et déjà organisées pour être lues par un programme, un logiciel à règles fixes suffit. L'IA devient utile quand les informations sont éparpillées, manquantes ou contradictoires.

Les erreurs que les agents font encore

Le plus souvent, l'agent choisit la bonne décision. La difficulté est de citer ses sources sans erreur.



POURQUOI LES RÉSULTATS SONT MOINS BONS AUX ACHATS

Pour la revue Achats, les agents utilisant Luna ou DeepSeek prennent la bonne décision dans 96 à 97 % des cas. Pourtant, seules 33 à 43 % des revues passent tous les tests : les agents citent un tableur que le programme de notation n'accepte pas comme source.

TRÈS PEU DE PROJETS APPROUVÉS À TORT

Sur environ 5 100 revues, aucun projet non conforme n'a été approuvé à tort avec Gemini. Il y a eu une telle erreur avec Luna et une avec DeepSeek.

LES RÉSULTATS CHANGENT SELON LA REVUE

Luna obtient de meilleurs résultats que Gemini en Architecture technique et en Préparation à l'exploitation, mais de moins bons résultats aux Achats et en Validation finale. Aucun modèle ne devance les autres dans tous les domaines.

Dans ces tests, les agents prennent généralement la bonne décision. Il reste surtout à mieux retrouver et recopier les passages qui justifient leurs conclusions.

Deux exemples concrets

Dans le premier cas, la revue réussit tous les tests. Dans le second, la décision est bonne, mais une citation incorrecte fait échouer la revue.

PROJET FALCON · PORTAIL RH · TEST RÉUSSI

- Lors de la revue Préparation à l'exploitation, l'agent découvre que le document expliquant comment exploiter le système, le runbook, est encore à l'état de brouillon.
- Il demande de terminer ce document et cite le passage qui montre qu'il n'est pas finalisé.
- Le système autorise l'agent à approuver le projet sous conditions.
- L'agent répond : APPROUVÉ SOUS CONDITIONS.

C'est bien la réponse prévue pour ce dossier. Chacun des trois modèles a permis à l'agent de réussir les tests des cinq revues de Falcon.

PROJET MERIDIAN · BONNE DÉCISION, CITATION INCORRECTE

- L'agent repère deux problèmes :
- la passerelle d'API n'est pas en place ;
- le temps de réponse de l'application est de 55 ms, alors qu'il ne doit pas dépasser 20 ms.
- Il demande de corriger ces deux problèmes et classe le projet À CORRIGER.
- Sa décision est correcte, mais il inverse l'ordre de deux informations dans une citation.
- Le test exige de reprendre le texte mot pour mot. La revue est donc considérée comme échouée.

L'agent a pourtant pris la bonne décision, repéré les bons problèmes et demandé les bonnes corrections. La revue échoue uniquement parce qu'une citation n'est pas exacte.

Dans cet exemple, le raisonnement de l'agent n'est pas en cause. Il faudrait relier automatiquement chaque problème au passage exact qui permet de le constater.

L’agent réussit-il encore quand on refait le test ?

Sur le même dossier, l’agent peut réussir un essai et échouer au suivant.

Gemini 3.8 Flash

9 sur 15

projets dont toutes les revues ont réussi lors de chacun des trois essais

35 sur 45

essais où toutes les revues du projet ont réussi

96 %

des revues ont réussi, tous nouveaux essais confondus

GPT-5.6 Luna

3 sur 15

projets dont toutes les revues ont réussi lors de chacun des trois essais

19 sur 45

essais où toutes les revues du projet ont réussi

83 %

des revues ont réussi, tous nouveaux essais confondus

DeepSeek v4.1 Flash

0 sur 15

projets dont toutes les revues ont réussi lors de chacun des trois essais

11 sur 45

essais où toutes les revues du projet ont réussi

73 %

des revues ont réussi, tous nouveaux essais confondus

L’AGENT DEMANDE L’AUTORISATION D’APPROUVER

L’agent peut demander à approuver un projet malgré un problème, si les règles permettent d’en accepter le risque sous certaines conditions. Un contrôle automatique vérifie ses autorisations, puis accepte ou refuse sa demande.

	Gemini	Luna	DeepSeek
Demandes envoyées pour approuver un projet sous conditions	**864**	**394**	**216**
Demandes refusées par le contrôle automatique	**466**	**40**	**14**
Approbations sous conditions retenues dans les décisions finales	**398**	**353**	**202**

Avec Gemini, l’agent demande plus souvent à approuver sous conditions. C’est aussi avec ce modèle que le système refuse le plus de demandes. Le contrôle automatique empêche l’agent de prendre une décision qui dépasse ses autorisations.

Un seul essai réussi ne prouve pas que l’agent est fiable. Avant de l’utiliser en entreprise, il faut refaire les tests pour vérifier qu’il reste fiable et qu’il ne peut pas prendre de décisions non autorisées.

Une information manquante peut changer la décision

Deux projets peuvent avoir les mêmes documents et demander des décisions opposées. Une seule information officielle, conservée ailleurs, fait la différence.

PROJET A

26 documents Word
exactement les mêmes que pour l'autre projet

Dans le tableur officiel de l'entreprise :
le fournisseur a été contrôlé

LA DÉCISION À PRENDRE

APPROUVÉ (GO)

=

des documents identiques

≠

une information qui change

≠

deux décisions opposées

PROJET B

26 documents Word
exactement les mêmes que pour l'autre projet

Dans le tableur officiel de l'entreprise :
le fournisseur n'a pas été contrôlé

LA DÉCISION À PRENDRE

À CORRIGER (REWORK)

Avec les seuls documents Word, ni un humain ni un agent IA ne peut savoir lequel des deux projets doit être approuvé. Il manque une information, pas une meilleure capacité de raisonnement.

L'agent doit donc pouvoir consulter les informations officielles de l'entreprise, même lorsqu'elles ne figurent pas dans le dossier. S'il ne peut pas les consulter, il doit pouvoir répondre : « Il me manque une information pour décider. »

Combien de travail reste-t-il aux humains ?

Le nombre de cas automatisés ne suffit pas : il faut mesurer le temps de travail qui reste aux équipes.

COMMENT CALCULER LE TRAVAIL HUMAIN RESTANT

Cas complexes : nombre de cas encore confiés aux humains × temps nécessaire pour chacun.

+ Vérifications : temps consacré aux signatures, aux contrôles sur un échantillon de dossiers et aux validations qui restent humaines.

+ Corrections : temps passé à repérer et à corriger les erreurs de l'agent, y compris leurs conséquences sur les revues suivantes.

+ Maintenance : temps passé à mettre à jour les connexions aux autres systèmes, les règles et les cas de test, sans oublier l'assistance du fournisseur.

Confier la même tâche à un autre service ne fait pas gagner de temps à l'entreprise. Ces heures doivent toujours être comptées.

LES CAS LAISSÉS AUX HUMAINS PRENNENT PLUS DE TEMPS

Si l'IA traite les 80 % de cas les plus simples, les 20 % restants peuvent encore occuper 40 % du temps de travail, voire davantage, parce qu'ils sont plus difficiles.

CORRIGER LES ERREURS PREND AUSSI DU TEMPS

Une mauvaise décision ajoute du travail : il faut d'abord la repérer, puis la corriger. Une erreur grave qui passe inaperçue reste inacceptable, même si cela arrive rarement.

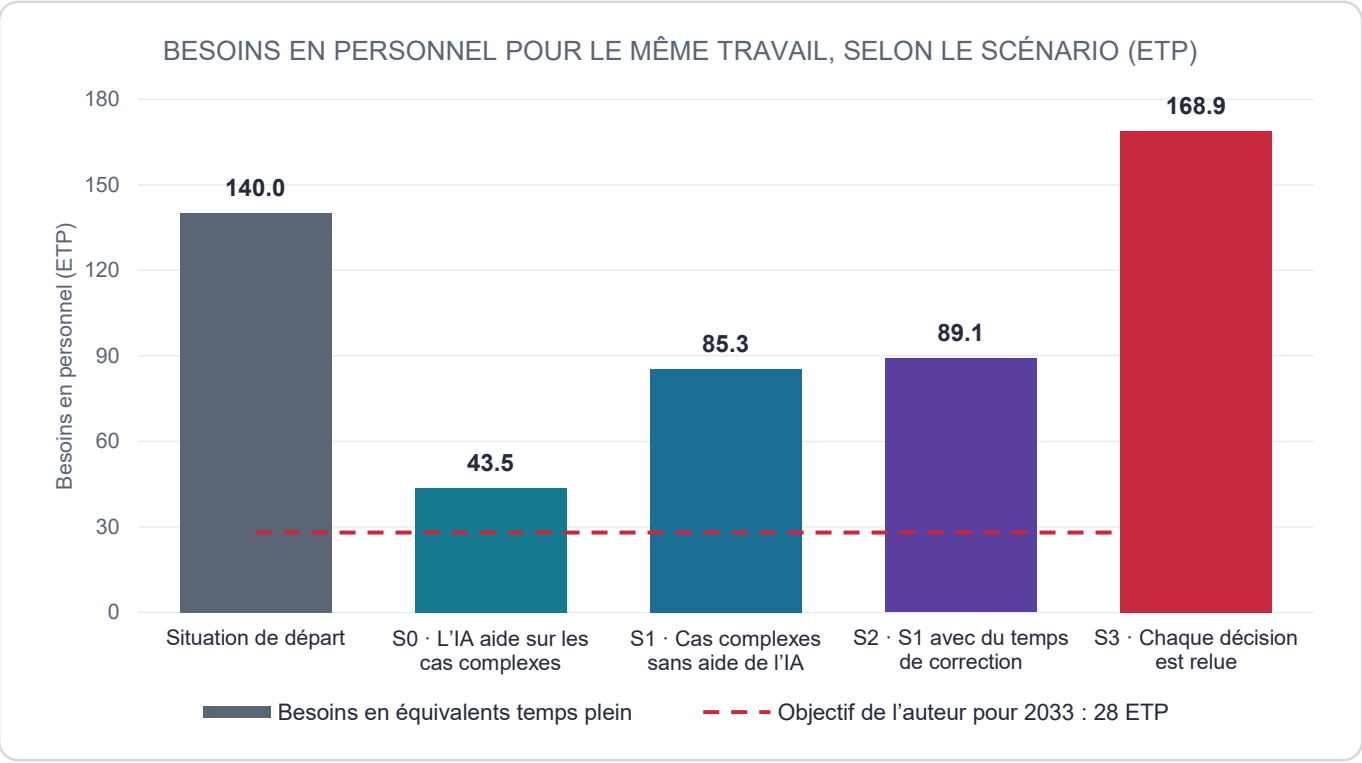
LES MESURES À RELEVER

Il faut compter les dossiers traités, les heures travaillées avant et après, les dossiers renvoyés aux humains, ainsi que le temps de correction et de maintenance. La comparaison doit porter sur des dossiers de même nature avant et après l'arrivée de l'IA.

Un bon taux de réussite aux tests ne dit pas combien d'heures l'IA fait gagner. Pour le savoir, il faut mesurer le travail qui reste aux humains.

Le même agent, mais des besoins humains très différents

Dans cet exemple, le travail représente aujourd'hui 140 équivalents temps plein (ETP), soit la charge de travail de 140 personnes à temps plein. Avec le même agent IA, les besoins vont de 44 à 169 ETP selon la façon d'organiser le travail.



S0 · L'IA aide aussi à traiter les cas complexes — **44 ETP · −69 %**

L'IA permet aux spécialistes de traiter les cas complexes en deux fois moins de temps. Les vérifications humaines restent limitées. Ce scénario optimiste ne compte pas le temps de maintenance.

S1 · Les humains traitent seuls les cas complexes — **85 ETP · −39 %**

Le traitement des cas complexes prend autant de temps aux humains qu'avant. Chaque équipe affecte une personne au fonctionnement de la plateforme.

S2 · On ajoute le temps de correction — **89 ETP · −36 %**

On reprend le scénario S1 et on ajoute 30 minutes de correction par cas.

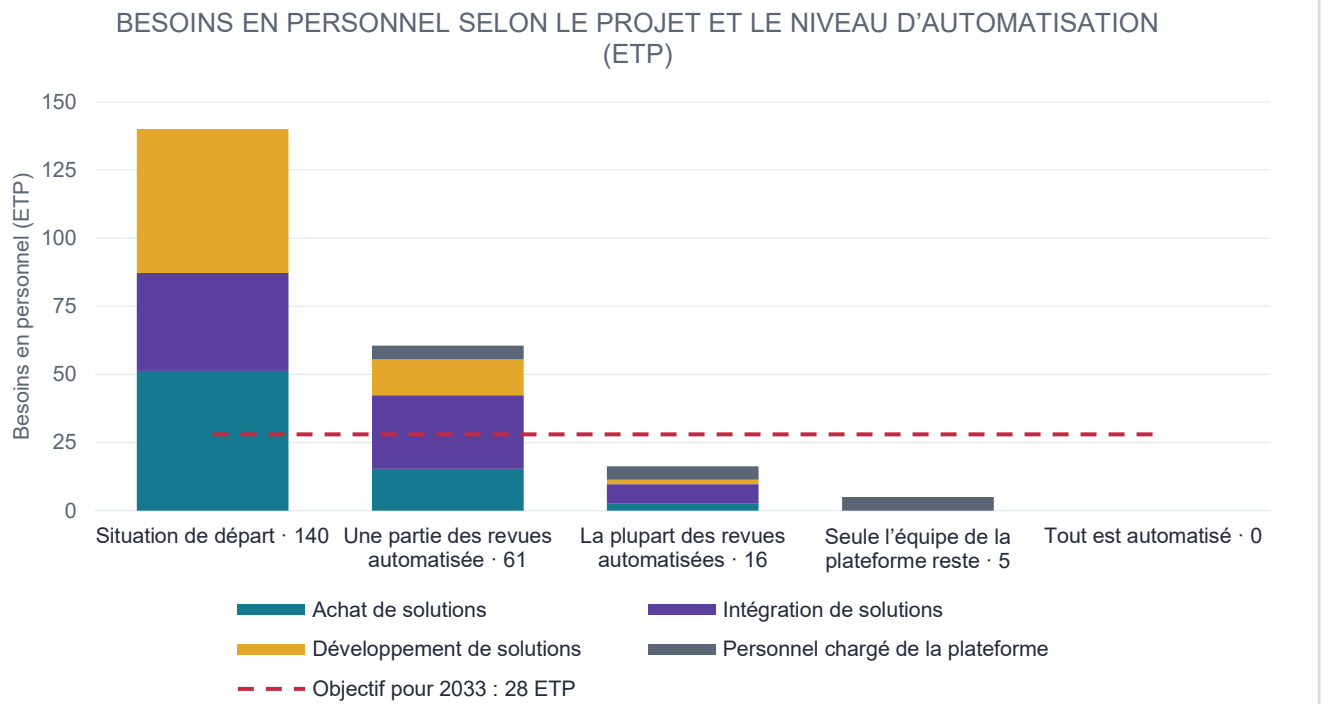
S3 · Un humain relit chaque décision de l'agent — **169 ETP · +21 %**

Chaque décision est relue par un humain. Il faut aussi compter les corrections et la maintenance. Il faut alors plus de travail humain qu'au départ.

Dans ces scénarios, le modèle d'IA est toujours le même. C'est l'organisation du travail humain qui change le résultat.

Jusqu'où l'automatisation pourrait-elle aller ?

Cinq scénarios théoriques montrent comment le travail humain pourrait passer de 140 ETP à zéro. Ils ne disent pas quand cela arriverait.



CE QUE DÉCRIT CHAQUE SCÉNARIO

Automatisation partielle : les humains réalisent encore certaines revues, surtout pour les projets d'intégration. Cinq personnes assurent le fonctionnement de la plateforme.
Automatisation majoritaire : selon le type de projet, les humains ne font plus que 3 à 20 % du travail initial.
Exploitation seule : toutes les revues sont automatisées, mais cinq personnes font encore fonctionner la plateforme.
Automatisation complète : le fonctionnement de la plateforme est lui aussi automatisé.

LES PERSONNES NÉCESSAIRES POUR FAIRE FONCTIONNER LA PLATEFORME

Même en supprimant 80 % du travail de revue, il reste 33 ETP au total si cinq personnes doivent faire fonctionner la plateforme, et non 28. Avec cinq personnes pour la plateforme, il faudrait réduire d'environ 84 % le temps humain consacré aux revues pour atteindre 28 ETP au total. Avec une équipe de quinze personnes, il faudrait le réduire d'environ 91 %.

CE GRAPHIQUE NE DONNE PAS DE CALENDRIER

Chaque barre montre le travail humain qui resterait à un niveau donné d'automatisation. La dernière barre illustre une possibilité théorique. Elle ne prédit pas l'avenir.

Les contrôles prévus restent obligatoires. Dans ces scénarios, il faut simplement moins de travail humain pour les effectuer.

Quand cela pourrait-il devenir possible ?

Il faut que l'agent sache mener des tâches plus longues, fasse très peu d'erreurs, réussisse des tests rigoureux et obtienne le droit de décider.

1 · RÉUSSIR DES TÂCHES PLUS LONGUES

D'après les chiffres cités dans la version longue, les meilleurs agents réussissent de façon fiable, en mai 2026, des tâches qui demandent environ trois heures à un humain. L'étude estime qu'une revue complète demande environ 24 heures de travail humain. Si les agents continuent à doubler tous les sept mois la durée des tâches qu'ils savent traiter, ils pourraient atteindre l'équivalent de 24 heures de travail humain début 2028.

2 · FAIRE BEAUCOUP MOINS D'ERREURS

Réussir la plupart du temps ne suffit pas. Dans ce scénario, il faut passer d'environ une erreur sur cinq cas à moins d'une sur mille.

3 · PROUVER SA FIABILITÉ PAR DES TESTS

Les spécialistes doivent préparer des cas de test, puis vérifier que l'agent les réussit. Le scénario prévoit environ un an pour cette étape, à titre d'exemple.

4 · DONNER À L'AGENT LE DROIT DE DÉCIDER

L'entreprise doit préciser ce que l'agent peut approuver, lui donner les autorisations nécessaires et obtenir l'accord du service juridique. Le scénario prévoit un an de plus, mais les obligations légales peuvent rallonger ce délai.

QUAND L'AGENT POURRAIT-IL RÉALISER UNE REVUE SANS QU'UN HUMAIN LA RELISE À CHAQUE FOIS ?

Selon les hypothèses choisies, une revue pourrait être confiée à l'IA sans relecture humaine systématique entre 2031 et 2034. Ces dates illustrent des scénarios possibles, pas des prévisions.

L’hypothèse : confier une revue entière à l’IA

À terme, l’agent pourrait traiter aussi les dossiers complexes et prendre la décision finale. La supervision de la plateforme pourrait elle aussi être automatisée.

CE QUE SUPPOSE L’HYPOTHÈSE

- L’hypothèse est qu’un agent IA pourrait effectuer chaque revue du début à la fin.
- L’IA pourrait aussi surveiller la plateforme, traiter les cas particuliers et mettre à jour les règles.
- Aucune étape de contrôle ne disparaît.
- L’agent IA réalise les contrôles à la place du spécialiste.
- Si l’automatisation réduit réellement le travail au lieu de le transférer, il pourrait ne rester qu’une petite équipe pour la plateforme, puis éventuellement plus aucun travail humain.

Cette hypothèse ne tient pas si une revue ou une signature doit rester humaine, si le service prétendument automatisé cache du travail humain, ou si l’équipe nécessaire à la plateforme ne diminue jamais.

SIX POINTS À OBSERVER POUR TESTER L’HYPOTHÈSE

1 · Les revues devraient être plus faciles à automatiser quand les informations sont claires, les règles écrites et les résultats vérifiables.

2 · Transmettre un dossier bien organisé d’une revue à l’autre devrait éviter de refaire le même travail.

3 · Les dossiers laissés aux humains devraient souvent être les plus longs à traiter.

4 · Pour estimer les heures gagnées, il faut compter les dossiers complexes, les corrections et la maintenance. Automatiser 80 % des cas ne fait pas forcément gagner 80 % du temps.

5 · Les entreprises devraient d’abord laisser l’IA préparer les décisions, puis lui donner le droit d’en prendre certaines.

6 · L’automatisation ne réduit le travail humain que si elle fait gagner plus d’heures qu’elle n’en ajoute.

L’expérience ne prouve pas cette évolution. L’étude propose une hypothèse, six points à suivre et des situations qui montreraient qu’elle ne se vérifie pas.

L’hypothèse pour 2033 : au moins 80 % d’ETP en moins

Pour confirmer l’hypothèse, il faudra traiter autant de dossiers, aussi bien et aussi vite, avec au moins 80 % d’ETP en moins.

−80 %

d’ETP pour les mêmes revues du DGF au 20 septembre 2033, par rapport au travail mesuré pendant l’année terminée le 20 septembre 2026.

Dans l’exemple de départ, cela revient à passer de 140 ETP à 28 ou moins.

TOUT LE TRAVAIL HUMAIN À COMPTER

Il faut compter toutes les heures : examen des dossiers, cas complexes, vérifications, corrections, maintenance, aide des fournisseurs et supervision par les responsables.

CE QUI MONTRERAIT QUE L’HYPOTHÈSE EST FAUSSE

L’hypothèse échoue s’il reste plus de 20 % du travail humain initial, si la qualité baisse ou si les délais augmentent. Elle échoue aussi si le calcul exclut les cas difficiles ou oublie les nouveaux besoins d’assistance.

QUAND ON NE PEUT PAS CONCLURE

Sans mesure des heures avant et après, impossible de savoir si l’hypothèse est confirmée ou non.

CE QUI NE PROUVE PAS UNE RÉUSSITE

Faire faire les mêmes heures à des prestataires, laisser passer des défauts ou renommer « supervision IA » une équipe qui refait les revues à la main ne compte pas comme une réussite.

L’étude précise ce qu’il faut mesurer et comment le calculer, sans laisser de côté le travail qui reste aux humains.

Ce que l'IA peut faire, et ce qui reste aux humains

Un agent IA peut prendre en charge certaines tâches. L'agent effectue le travail, mais la responsabilité reste à l'entreprise et aux personnes qu'elle a désignées.

LES TÂCHES QUE L'AGENT PEUT EFFECTUER

- Examiner tous les documents du dossier
- Rendre une décision et expliquer pourquoi
- Appliquer les règles et vérifier les seuils fixés
- Indiquer précisément où il a trouvé chaque problème
- Transmettre à la revue suivante les conditions qui restent à remplir
- Approuver sous conditions, mais seulement s'il en a le droit

LE TRAVAIL QUI RESTE AUX HUMAINS ET QU'IL FAUT COMPTER

- Définir et valider les règles
- Décider dans les cas où l'agent n'a pas l'autorisation de le faire
- Apposer les signatures humaines exigées par la loi ou l'entreprise
- Vérifier un échantillon des dossiers traités
- Faire fonctionner la plateforme et en assurer la maintenance
- Réaliser les actions demandées par la revue tant qu'elles ne sont pas automatisées
- Renommer les postes ne supprime pas le travail

Appeler les personnes « superviseurs » ne change pas le nombre d'heures qu'elles travaillent.

Moins de travail peut conduire à supprimer des postes. L'entreprise peut aussi garder la même équipe et lui faire traiter davantage de dossiers.

Renvoyer tous les dossiers à un humain n'automatise pas la revue

Si l'agent confie toutes les décisions à un humain, il limite les risques qu'il prend, mais ne remplace aucune revue humaine.

Les spécialistes se consacrent davantage aux règles

Ils passent moins de temps sur chaque dossier et plus de temps à définir, tester et améliorer les règles de décision.

L'entreprise reste tenue de faire réaliser les contrôles obligatoires et reste responsable des décisions. Ce qui change, c'est la répartition des tâches entre les humains et les agents IA.

Et dans les autres métiers soumis à des règles ?

On y retrouve le même travail : lire un dossier, appliquer des règles, prendre une décision et l'expliquer.

DES DOSSIERS À EXAMINER, DES DÉCISIONS À PRENDRE

Exemples : souscrire une assurance, traiter un sinistre ou demander un prêt. On commence par examiner un dossier. On aboutit à un tarif, à un paiement, à un accord ou à un refus.

- **Dans l'assurance :** souscripteurs et gestionnaires de sinistres.
- **Dans la banque :** chargés de crédit et analystes crédit.
- Autres exemples : prêts immobiliers, déclarations en douane et suivi des effets indésirables des médicaments.

La version longue envisage cette possibilité pour d'autres métiers. L'expérience sur le DGF n'a pas testé ces métiers.

Allianz · l'IA prépare les dossiers, un humain autorise le paiement

Dans l'exemple cité en Australie, sept agents IA traitent de petits sinistres. Le paiement prend quelques heures au lieu de plusieurs jours, mais chaque paiement doit encore être approuvé par un gestionnaire humain. Le délai est plus court, mais cela ne dit pas combien d'heures de travail humain ont été gagnées.

AIG · l'IA prépare l'analyse, le spécialiste décide

Chez AIG, des agents IA lisent les dossiers, évaluent les risques, comparent les prix et rédigent une synthèse. C'est toujours l'analyste humain qui prend la décision.

Upstart · la plupart des prêts financés sont approuvés sans revue humaine

Plus de 90 % des prêts financés par Upstart sont approuvés sans revue humaine. En revanche, environ un quart de l'ensemble des demandes est renvoyé à une personne. Le premier chiffre concerne les prêts financés. Le second concerne toutes les demandes. On ne peut donc pas les comparer directement.

POURQUOI CES MÉTIERS SONT PLUS DIFFICILES À AUTOMATISER

- **Ces décisions touchent directement les personnes, ce qui rend l'automatisation de ces métiers plus délicate.** Certains de ces usages de l'IA sont classés « à haut risque », et le droit européen encadre les décisions entièrement automatisées. Il faut aussi adapter l'automatisation aux règles et au fonctionnement de chaque secteur.

Ces exemples montrent que certaines tâches simples sont déjà automatisées. Les dossiers complexes restent souvent confiés aux humains. L'hypothèse est que l'IA pourra progressivement traiter aussi ces dossiers plus difficiles.

Ce qu'il faut retenir

Les contrôles prévus restent obligatoires. L'agent IA peut effectuer certaines revues. Il faut compter tout le travail qui reste aux humains.

Les FDE rendent l'IA utilisable dans l'entreprise

Les FDE donnent aux agents IA accès aux documents, aux règles, aux applications et aux informations officielles dont ils ont besoin. L'article présente les revues du DGF comme un premier usage possible de cette approche.

Des agents IA peuvent examiner les dossiers

Les revues du DGF reviennent régulièrement, suivent des règles écrites et donnent des résultats vérifiables. Un agent peut les réaliser s'il a accès aux informations officielles nécessaires aux contrôles.

Un même agent ne réussit pas à tous les essais

Dans les tests, les agents ont souvent pris la bonne décision. Ils doivent encore mieux citer leurs sources et obtenir des résultats réguliers d'un essai à l'autre.

Automatiser davantage ne dit pas combien de postes disparaîtront

Dans les scénarios étudiés, le même agent peut laisser un besoin de 44 à 169 ETP, selon la manière d'organiser le travail. Le vrai point à mesurer est le temps qui reste aux humains, pas seulement le nombre de cas automatisés.

Dans l'expérience, l'agent réussit une revue s'il prend la bonne décision, repère les bons problèmes, demande les corrections nécessaires, cite les bonnes sources et respecte ses autorisations.
La version longue envisage qu'à terme, même la supervision de la plateforme puisse être automatisée.
L'étude ne mesure pas les économies dans une entreprise réelle, ne compare pas directement les agents à des spécialistes humains et ne montre pas quels secteurs adopteront l'IA en premier.

Consulter l'étude et vérifier les résultats

On peut vérifier les résultats publiés sans solliciter à nouveau les modèles d'IA.

LES DEUX VERSIONS DE L'ÉTUDE ET LES DONNÉES

L'article de 28 pages présente l'idée, l'expérience et le calcul des effets sur l'emploi. Il est disponible sur arXiv : arxiv.org/abs/2609.29345.
github.com/jeremy1392/dgf-agentic-bench/blob/main/paper2/From_Governance_Reviews_to_Task_Substitution.pdf
La version longue de 66 pages détaille l'hypothèse, les délais possibles, les objections et les scénarios à long terme.
github.com/jeremy1392/dgf-agentic-bench/blob/main/paper/The_Last_Human_Gate.pdf
Les données publiées comprennent les 300 dossiers, les réponses des agents, les scores, les coûts et les essais répétés. Elles incluent aussi la version du code utilisée et des empreintes SHA-256 pour vérifier que les fichiers n'ont pas changé.
Pour refaire les calculs, téléchargez les données, vérifiez les fichiers grâce aux empreintes SHA-256, puis relancez la notation et les statistiques.
Il n'est pas nécessaire de faire de nouveaux appels aux modèles d'IA.

LES TRAVAUX CITÉS DANS L'ARTICLE

Acemoglu & Restrepo (2019). Automation and new tasks. Journal of Economic Perspectives.
Bainbridge (1983). Ironies of automation. Automatica.
Calvanese et al. (2026). Agentic business process management. Information Systems.
Cooper (1990). Stage-gate systems. Business Horizons.
Demirer, Horton, Immorlica, Lucier & Shahidi (2026). Chaining tasks, redefining work. NBER.
Drouin et al. (2024). WorkArena. ICML.
Gans & Goldfarb (2026). O-Ring automation. NBER.
Gao, Yen, Yu & Chen (2023). Enabling LLMs to generate text with citations. EMNLP.
Jha et al. (2025). ITBench. arXiv:2502.05352.
Kapoor, Stroebl, Siegel, Nadgir & Narayanan (2024). AI agents that matter. arXiv:2407.01502.
Ly, Maggi, Montali, Rinderle-Ma & van der Aalst (2015). Compliance monitoring in business processes.
Palantir Technologies (2026). Forward Deployed Software Engineer, fiche de poste.
Rashkin et al. (2023). Measuring attribution in natural language generation models.
Yao, Shinn, Razavi & Narasimhan (2024). τ-bench. arXiv:2406.12045.

Les tests portent sur les modèles proposés par OpenRouter les 22 et 23 septembre 2026. L'expérience n'utilise aucune donnée provenant d'une entreprise réelle.

Merci !

Place aux questions

Lisez l’article, refaites l’expérience et vérifiez l’hypothèse : au moins 80 % d’ETP en moins en 2033.

https://www.linkedin.com/in/jcanale13/

Jeremy Canale · AI Applied Security Architect

contact@jeremycanale.com · https://www.jeremycanale.com

Article : arxiv.org/abs/2609.29345 · Étude complète et données : github.com/jeremy1392/dgf-agentic-bench